# Intelligent Information Retrieval and Recommender System Framework

Sitalakshmi Venkatraman, *Senior Member*, *IACSIT* and Sadhana J. Kamatkar

***Abstract*—With huge growth in enterprise data, intelligent information retrieval methods have gained research focus. This paper addresses the difficulty of retrieving *relevant, pertinent,* and *novel* information for a large system that involves fusion of data in different formats such as, text, barcode, and images. We propose a framework to combine an intelligent image retrieval and intelligent information retrieval (IIR) along with the user profile learning to develop a recommender system. We demonstrate the application of our proposed framework in a real-life situation.**

***Index Terms*—Information retrieval, data mining, user profile, recommender system**

## I. Introduction

Recently, we witness exponential increase in the amount of information being produced. Effective decision making based on such huge amounts of data can be achieved only if useful knowledge is extracted automatically from them [1]. Hence, intelligent information retrieval (IIR) methods and policies are warranted for an efficient assimilation of such information leading to timely and productive decision making. IIR using suitable data mining of user profiles provides the means to categorise data. This facilitate in the reuse and organisation of data for synthesising knowledge required for recommender systems [2] [3].

Ideally, the intelligence aspect of data mining should be adopted to provide decision support as it attempts to discover patterns, trends and correlations hidden in data to help in making effective decisions [4] [5]. However, traditional data mining techniques are not capable of combining text, image and user profile data to retrieve pertinent information for providing good recommender systems for effective and timely decision making.

This paper proposes a new intelligent information retrieval (IIR) approach to develop a recommender system framework for processing large data sets of multiple formats, including text, barcode, image. The recommender system is based on user profile learning by combining 'data relevance' from multiple sources that facilitate in arriving at a reduced data set intelligently. We apply our proposed framework in the context of a real-life 'Image Retrieval System' developed for a large data processing center within a university setting. Through the real-life application, we demonstrate the use of a IIR framework that leads to a recommender system based on data mining of the user profiles and relevance feedback.

Manuscript received September 17, 2012; revised October 29, 2012.

S. Venkatraman is with the School of Science, Information Technology and Engineering, University of Ballarat, PO Box 663 VIC 3353, Australia (e-mail: s.venkatraman@ballarat.edu.au).

S. J. Kamatkar is with Central Computing Facility at University of Mumbai, India - 400032 (e-mail: sjk@mu.ac.in).

## II. Literature Review

The information retrieval (IR) process of data mining requires extraction of valid patterns and relationships in very large data sets automatically [6]-[7]. It is usually portrayed like a "voyage into the unknown", and hence requires the use of techniques from AI and statistics, such as machine learning, pattern recognition, classification, and visualization [8]-[9]. In this modern digital age with exponential growth in business data, literature studies have proved that for achieving real-time data mining applications, the speed of analysis could be improved by adopting highly efficient information retrieval (IR) methods [10]-[11]. In a nutshell, a general definition of IR is, "Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information" [12], and the IR methods employed depend on the organisation and type of data [13]-[14]. Since 1950s, the primary focus of IR has been on text and documents, and more recently, with applications increasingly involve new media e.g., video, photos, music, and speech, traditional IR techniques are seeking contemporary approaches that are more intelligent to search and evaluate data in these new media. It is important to focus attention on the most relevant variables such as user-based information that could be employed for information filtering and intelligent retrieval in data mining of the Web [15]-[16] [17]. Information filtering could be performed by gathering patterns of user interaction with the systems and constructing user profiles using various learning techniques such as, genetic algorithms, probabilistic model of clustering and relevance-based ontology [18]-[20].

In dealing with image-based information retrieval, previous research studies have only employed classical information retrieval [21]-[23] or traditional data mining techniques [24]-[26] for large data, and they are not capable of combining text, image and user profile data to retrieve relevant information for a faster filtering process. In this paper a novel approach to combine contemporary IR techniques that is based on user profile constructed from multiple source of 'data relevance' along with data mining learning algorithms in arriving at a recommender system. This paper aims to address the gap in literature by proposing an intelligent information retrieval (IIR) for the development of a recommender system.

DOI: 10.7763/IJFCC.2013.V2.127

## III. Intelligent Information Retrieval Approach

We propose an intelligent information retrieval (IIR) approach that is designed to integrate two disparate methods, namely information filtering and data mining. Simple IR could be time consuming and may not be achievable without manual interventions for data sets that involve different media such as video, audio, images and documents [27]. Our Intelligent IR takes into account the meaning of the words used in the query, their relationships such as the order of words in the query, and thereby establishes the relevance. It is also designed to adapt the query based on the user's direct and indirect profile contexts and relevance feedback [28]-[30]. Our model of IIR makes use of information utility and relevance.

Though utility and relevance are important for all IR operations, measuring them and using them intelligently is important [31]. Utility might be measured in monetary terms: "How much is it worth to the user to have found this document?" "How much did we save by finding this software?" In the literature, the term "relevance" is used imprecisely; it can mean utility or topical relevance or pertinence. Many IR systems focus on finding topically relevant documents, leaving further selection to the user. Relevance is a matter of degree; some documents are highly relevant and indispensable for the user's tasks; others contribute just a little bit. From relevance assessments we can compute measures of retrieval performance such as

*Recall*: How good is the system at finding relevant documents?

*Discrimination*: How good is the system at rejecting irrelevant documents?

*Precision*: Depends on discrimination, recall, and the number of relevant documents.

Evaluation studies commonly use recall and precision or a combination. With low precision, the user must look at several irrelevant documents for every relevant document found. More sophisticated measures consider the gain from a relevant document and the expense incurred by having to examine an irrelevant document [32]-[33]. For example, many relevant documents that merely duplicate the same information just waste the user's time, so retrieving fewer relevant documents would be better.

In this paper, we propose an intelligent IR approach considering three main utility attributes, *relevant, pertinent and novel*, for retrieving a document from a large data base. A document is topically relevant for a particular situation, context, query, or task if it contains information that either directly answers the query or can be used, possibly in combination with other information, to derive an answer or perform the task. It is pertinent with respect to a user with a given purpose if, in addition, it gives just the information needed; is compatible with the user's background and cognitive style so s/he can apply the information gained and is authoritative. It is novel if it adds to the user's knowledge i.e. finding unknown things which is the part of data mining. In this paper, we propose an intelligent IR approach to construct a user model through gaining relevant user feedback, which can significantly arrive at a smaller set of ranked documents that are relevant to the user's interests or search intent. A learning technique could then be adopted to arrive at a user profile based on how well the documents are topically relevant and pertinent.

## IV. The Need for a Recommender System

Many applications make use of barcode and Optical Mark Recognition (OMR) systems to aid in automated information processing. Typically, we find their use in many university settings, where examination question papers and answer booklets adopt such systems and the answer booklets are captured by the system as images. One of the main reasons for this reform towards the use of OMR is firstly to protect the identity of the student writing the exam and the seat number of the student, before the answer booklet goes for evaluation to the examiner. Secondly, and more importantly, OMR systems are expected to aid in automate the process of answering the variety of queries raised that relate to the students' exam results after evaluation. Such queries on the answer booklets require intelligent information retrieval and a recommender system.

One of the universities we considered for proposing our IIR approach was University of Mumbai. This university conducts the examinations twice a year, and they are referred as First Half i.e. examinations conducted in April-May and Second Half i.e. examinations conducted in October-November. The total numbers of students enrolled in First Half 2011 were around 300,000 (Three Hundred Thousand). Thus the volume of data is very high. For example, the B.Com. Examination of First half 2011, had around 80,000 candidates who appeared for their final examinations. B.Com. has 64 subjects and from these subjects, student can opt for 7 subjects. Thus there are (80000 × 7 = 560,000 answer booklets) i.e. half million and sixty thousand records (answer books) for this one exam alone. Each answer booklet has 2 parts (560000 × 2 = 1,120,000), and hence there results in One Million Twenty Thousand images for this single exam alone.

In the barcode system, the first page of each answer booklet is divided into two parts – the first part contains the student information like seat number, subject code, centre code, etc. along with unique bar code. The second part contains examiner information such as marks awarded for each question, total marks scored, subject code, signature of examiner, bundle number; answer book number etc. along with unique bar code. This two parts are scanned and the images are stored in the database.

After declaration of examination results, students may apply to make a query of their marks obtained, such as, verification of marks, revaluation of marks, view their answer booklet, etc.

Due to huge number of answer booklets collected during every exam, and considering security concerns as well as cost, manpower, and other resources involved in scanning all the pages of the entire answer booklet, the management has decided to only scan the first page that provides two parts, namely student information and examination information. Fig. 1 gives a sample image capturing the first page of the exam booklet. All the answer booklets are stored physically in a storage location, and in order to answer any of the user queries, the answer booklet is manually retrieved from the

physical storage location. To achieve this, information relevant to the query had to be filtered and extracted from this huge data of images manually for identifying the correct storage location code for physical retrieval of the booklet. With growing number of student population and different subjects and degrees being offered, it is not practical to accomplish this manually for a timely answering of these queries. In addition to the problem of timely response to the queries, there are several drawbacks of the existing system. We adopted a systematic feedback and review of their manual and IT systems to reveal the drawbacks [32]-[33]. Some of the main drawbacks are that there is no system verification, and no recommendation given when there is missing information. It warrants a recommender system that can support keyword search and ad-hoc queries.

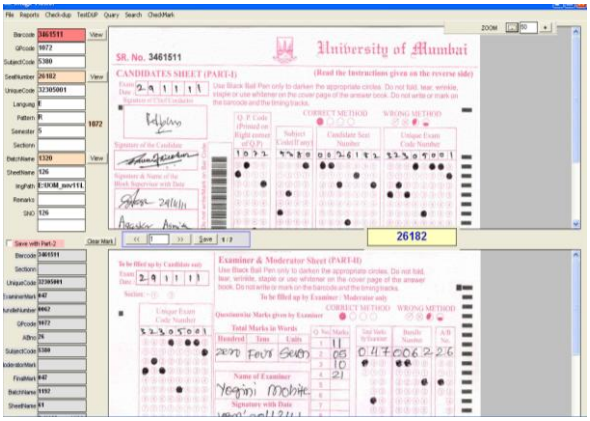

Fig. 1. Sample scanned exam booklet with student and exam information

## V. Proposed IIR and Recommender System

A real-life situation described earlier is a typical example where IT plays an important role in automating the processes for improving customer service and operations [34]-[35], and in particular where intelligent IR approaches are warranted. In this paper, we propose a novel approach that utilizes relevant and pertinent information in the intelligent information searching and filtering process, and provides ranking through data mining and a recommender system by constructing a user model. Our proposed system is designed to overcome the various drawbacks present in the real-life situation described in previous section. It consists of three main components, namely

Intelligent Image Retrieval Component
Intelligent IR Component
Recommender Component

### A. Intelligent Image Retrieval Component

This component consists of three main steps as described below. Each step includes system verification and validation procedures to maintain the integrity of the data stored in the system.

*Uploading of images* – This step uploads the images of all answer booklets consisting of First part (Student information) and Second part (Exam information). The images of First part have examination code, subject code, Bar code and other information of Student. While images of Second part have question wise and total marks given by Examiner or in some cases marks are also given by Moderator; examination code, subject code, bar code, bundle number, & answer book number.

*Linking of images* – Images of First part and Second part are linked using the Barcode. Also, they are linked by examination code and seat number so that retrieval can be fast. For example, a query to retrieve the image based on seat number or subject code (examination code) would be much faster with such associations established.

*Retrieving of images* – This step interfaces with the front-end of the system which communicates with the query module to retrieve the images. From these images the information about bundle number and answer booklet number can be used to retrieve the exact location in the storage for faster physical retrieval of the answer booklet. However, not all queries have the necessary inputs, in which case intelligent search and filtering of data is required. Hence, in such cases, it is necessary to interface with the recommender system based on data mining of relevance and user profile to have fast retrieval of required information.

This component is useful to many departments involved in facilitating the query process, such as the Physical Storage department, Photocopy Services department, Revaluation department, Student Services department, etc. For example, currently, a photocopy of the answer booklet is made whenever there is a view request made regarding the exam answer booklet.

### B. Intelligent IR Component

The intelligent IR component facilitates the recommender systems to focus on the user context-based recommender paradigm by using keywords and relevance criteria to arrive at the correct answer booklet from the database. This is achieved using the following processes:

- User interface to manage interaction with the user for query input and document output. relevance feedback. and visualization of results
- Keywords such as part of student name or subject name with relevance to some known data are intelligently processed using data mining approaches to form index words (tokens).
- Indexing constructs an inverted index of words resulting in document pointers.
- Searching retrieves documents that contain a given query token from the inverted index.
- Information integration and extraction to arrive at relevance. Relevance is a subjective judgment and includes the context, timeliness, authoritative and satisfaction level of a query.
- Associate ranking scores to all retrieved documents according to a relevance metric. The metric used here is based on data mining of item similarity (such as keywords) and user-item interaction (such as previous queries or user context). Similar approaches are found in literature [36][37].
- Query operations to transform the query in order to improve retrieval. We adopt query expansion using a thesaurus and query transformation using relevance feedback. While relevance-based probabilistic model for retrieval are adopted in some studies [31], we adopt relevance feedback with learning for ranking so that the ordering of answer booklets recommended could be

improved beyond what is possible with just relevance feedback alone [38]- [39].

- Information filtering (spam filtering) - this feature authenticates data and also looks for any spam that require to be categorised and removed.
- Automated document categorization - this feature is used to allocate an appropriate physical storage location for storing the answer booklet.
- Information clustering and routing - this process does the information clustering and routing for initiating processes pertaining to other departments such as Photocopy Services department, and Revaluation department.
- Recommending information using data mining of keywords, and combining with relevance feedback. These metrics are used to serve for intelligent filtering and recommendation of pertinent answer booklets to the user.

### C. Recommender Component

We introduce a recommender component in our proposed approach of intelligent IR so as to cater to even ad-hoc queries, keyword based search queries and incomplete queries to search the database for the relevant answer booklets. Recommender systems enhance the query interpretations by exploiting textual descriptions of the items to be recommended [40] and relevance rates given by users to infer a profile that is used to recommend items of interest [38]. In the proposed framework, the recommender component works on the user profiles and "image documents" (answer booklets) with the following key features:

Present topics/examinations that are of interest to the user

List the topics/examinations depending on the relevance such as, date, star recommendation or ratings.

Provide ranking scores to all retrieved image documents according to relevance metric.

Compare user's profile to some 'reference characteristics' to predict whether the user would be interested in an unseen item. We determine reference characteristics with content-based/ collaborative filtering approaches:

Information about the unseen item (content-based), and

User's social environment (collaborative filtering).

Our proposed recommender component is based on two main aspects: Building user profiles, and Learning user models.

#### 1) Building user profiles

Most recommender systems build a profile of user's interests, while our proposed recommender system, in addition, focuses on relevance and user context. This profile consists of two main types of information:

- user's preferences or interested in the item, and
- user's interaction history.

The system employs user's history as training data to create a user model. Predominantly, there are two main approaches:

- "Manual" recommendation approaches, where the user customization is done using a simple database matching process to find items that meet the specified criteria and recommend these to the users. However, this approach has major limitations such as, a) require additional effort from users to provide these set of criteria, b) it cannot cope with changes in user's interests and c) it does not provide a way to determine an order among recommending items.
- "Rule-based" recommendation approaches, where the system has rules to recommend other items based on user history. Our framework uses rule-based approach as it can capture common reasons for recommendations, which is highly effective in retrieving answer booklets.

Overall, the implementation of our proposed framework has improved the efficiency of the image-based information retrieval system. The average lead-time in data processing of a query, which was about a week, has been reduced to less than a minute's computer time with our proposed approach.

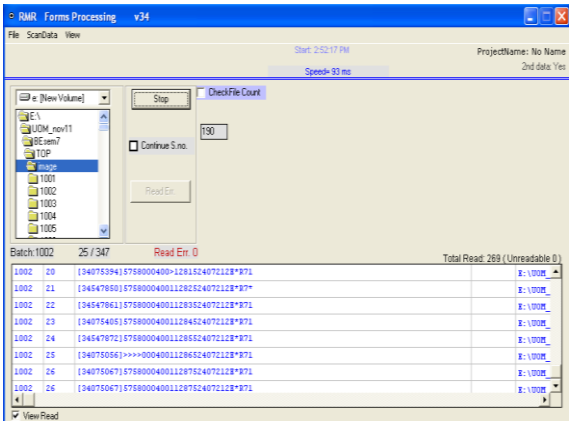


Fig. 2. Sample output listing a set of items based on relevance

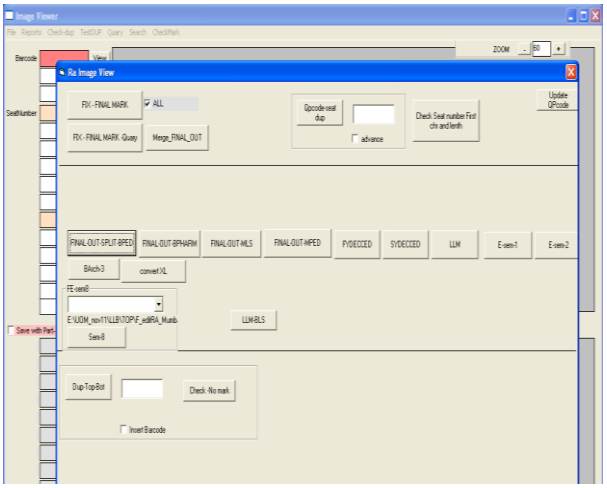


Fig. 3. Sample output showing fields of data extracted from the IR results

## VI. Conclusion

This paper discussed the importance of recommendation and personalization approaches and proposed an intelligent information retrieval and recommender system framework. The framework was implemented in automating the process of a physical search of the answer booklets of student examinations in a university setting. As the huge data comprises of many formats, including barcodes, student information, exam information, and image data, it has become mandatory to perform intelligent automation and to devise the methods of retrieving the information which is relevant for a topic/examination, pertinent with respect to a user profile and novel resulting in unknown user knowledge. The real-life implementation has demonstrated that learning a user model makes the recommender system highly effective.

## REFERENCES


[1] E. Turban and J. E. Aronson, *Decision Support Systems and Intelligent Systems*, Sixth Ed. New Jersey, Prentice Hall, 2001.

[2] K. E. Pearlson, *Managing and Using Information Systems: A Strategic Approach*, New York, Wiley, John Wiley and Sons, Inc, 2001.

[3] W. Frawley, G. P. Shapiro, and C. Matheus, "Knowledge Discovery in Databases: An Overview," *AI Magazine*, 1992.

[4] R Srikant and R Agrawal, "Mining sequential patterns: Generalizations and performance improvements," in *Proc. of the 5th International Conf. on Extending Database Technology*, France (March), 1996.

[5] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, Cambridge, Massachusetts, the MIT Press, 2001.

[6] R. Bradman and T. Anand, "The Process of Knowledge Discovery in Databases: A Human-Centered Approach," *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA: The AAAI Press/the MIT Press, pp. 37, 1996.

[7] U. Fayyad, G. P. Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA: The AAAI Press/The MIT Press, 1996.

[8] C. Westphal and T. Blaxton, *Data Mining Solutions - Methods and Tools for solving real world problems*, Wiley Computer Publishing, USA.

[9] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi, *Discovering Data Mining from Concept to Implementation*, Prentic Hall PTR, Inc. USA, 1998.

[10] R. B. Yates and B. R. Neto, *Modern information retrieval*, Reading, MA: Addison Wesley.

[11] H. Chu, *Information representation and retrieval in the digital age*, Medford, NJ: Information Today.

[12] G Salton, *Automatic Information Organization and Retrieval*, New York: McGraw-Hill, 1968.

[13] D. Soergel, *Organizing information: Principles of database and retrieval systems*, Orlando, FL: Academic Press, 1985.

[14] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[15] X. Zhou, Y. Li, Y. Xu, and R. Lau, "Relevance assessment of topic ontology," *in Proc. the fourth International Conference on Active Media Technology*, Relevance Assessment of Topic Ontology, 2006.

[16] S. J Kamatkar, "Information Retrieval in Web Mining for Knowledge Discovery," in *Proc. International Conference on Machine Learning and Computing*, 2011.

[17] P. Gawrysiak, "Information retrieval and the Internet," *PWII Information Systems Institute Seminars*, 1999

[18] D. Gibson, J. Kleinberg, and P. Raghavan, "Inferring Web communities from link topology," in *Proc. the 9th ACM Conference on Hypertext and Hypermedia*, 1998

[19] S. C. Cazella and L. O. C Alvares, "Modeling user's opinion relevance to recommending research papers," in *Proc. 10th International Conference on User Modeling (UM'05)*, Edinburgh, Scotland, 2005.

[20] Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs," *IEEE Trans. on Knowledge and Data Engineering*, vol. 18, no. 4, 2006, pp. 554-568

[21] D. C. Blair, *Language and representation in information retrieval*, Amsterdam: Elsevier Science, 1990.

[22] M. S. Silver, *Systems that Support Decision Making*, John Wiley, Chichester, 1991

[23] K. C. Laudon and J. P. Laudon, *Management Information Systems Organization and Technology*, Fourth Edition, Prentic-Hall of India, India, 1996.

[24] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proc. of 7th International World Wide Web Conference (WWW7). Computer Networks and ISDN Systems*, 1998.

[25] S. Brin and L. Page, "Anatomy of a large-scale hyper textual Web search engine," *WWW7 Conf. Proceedings*, 1998.

[26] R, Agarwal, "Data Mining," in *Proc. of International conference on Very Large Data Bases (VLDB)*, 1996.

[27] P. Buneman, S. Davidson, and D. Suciu, "Programming constrcuts for unstructured data," in *Proc. ICDT'95*, Gubbio, Italy,

[28] M. Balabanovic, S. Yoav, and Y. Yun, 1995 "An adaptive agent for automated web browsing," *Journal of Visual Communication and Image Representation*, vol. 6, no. 4.

[29] Z. Broder, S. C. Glassman, M. S. Manasse, and G Zweig, "Syntactic clustering of the web," in *Proc. of 6th International World Wide Web Conference*, 1997.

[30] C. Chang and C. Hsu, "Multi-engine search tool with clustering," in *Proc. of 6th International World Wide Web Conference*, 1997.

[31] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *Proc. the 24th annual international ACM SIGIR conference on Research and development in information retrieval ACM SIGIR*,1997.

[32] S. Venkatraman, "A Framework for ICT Security Policy Management," Esharenana E. A. (Ed.), *Frameworks for ICT Policy: Government, Social and Legal Issues*, IGI Global Publishers, USA, pp. 1-14, 2001.

[33] S. J. Kamatkar, *Computer and Applications: A Desktop Quick Reference*, Rushwin Publisher, India, 2003.

[34] S. J. Kamatkar, "Information Technology an important tool for Management in 21st century," in *Proc. of National Seminar organized by University of Mumbai*, India, 2001.

[35] S. J. Kamatkar, "Role of Information Technology(IT) in Globalisation Era," in *Proc. of International conference, organized by University of Mumbai*, India, 2002

[36] G. Pandey and J. Luxenburger, "Exploiting session context for information retrieval- a comparative study," *ECIR*, pp. 652-657. 2008.

[37] S. Verberne, H. V. Halteren, S. Raaijmakers, D. Theijssen, and L. Boves, "Learning to Rank for Why-Question Answering," *Information Retrieval*, 2010.

[38] P. L. Doughtie and K. Hofmann, "Learning to rank from relevance feedback for e-discovery," *ECIR*.

[39] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," *Information Retrieval*, vol. 13, no. 3, pp. 201-215, 2010.

[40] B. Xu, J. Bu, C. Chen, and D. Cai, "An Exploration of Improving Collaborative Recommender Systems via User-Item Subgroups," in *Proc. of International World Wide Web Conference*, 2012.


**Sitalakshmi Venkatraman** has obtained doctoral degree in Computer Science, from National Institute of Industrial Engineering, India in 1993 and MEd from University of Sheffield, UK in 2001. Prior to this, she had completed MSc in Mathematics in 1985 and MTech in Computer Science in 1987, both from Indian Institute of Technology, Madras, India.In the past 25 years, Sita's work experience involves both industry and academics. She has taught a variety of IT courses for tertiary institutions, in India, Singapore, New Zealand, and more recently in Australia since 2007. Currently, she is Senior Lecturer at the School of Science, Information Technology and Engineering, University of Ballarat, Australia. Her research supervisions have been in the areas of Mobile Usability, Biometrics, E-Security, E-Government and E-Health. At the University of Ballarat, she supervises research projects CIAO (Centre for Informatics and Applied Optimization). Sita has published seven book chapters and more than 70 research papers in internationally well-known refereed journals and conferences that include *Information Sciences*, *Journal of Artificial Intelligence in Engineering*, *International Journal of Business Information Systems*, and *Information Management & Computer Security*. She serves as Program Committee Member of several international conferences and Senior Member of professional societies and editorial board of three international journals

**Sadhana J. Kamatkar** has completed her higher education that includes M.Sc. in Statistics (1983), Diploma in Computer Management(1989), Ph.D. in Computer Science(2004), all from University of Mumbai, India. She became a Member (M) of IAENG in 2000 and a Senior Member (SM) in 2008, Sadhana has a total of 27 years of experience in IT field. Started career as Lecturer of Statistics to senior college students of M.S.G. College, Malegaon, Nasik, Maharashtra Subsequently, she worked in Hindustan Aeronautics Limited, Nasik for three years, and then in S.N.D.T. Women's University, Mumbai for six years. Presently working as I/c Director in Central Computing Facility at University of Mumbai, India. Sadhana has broad research interests that include Data ware Housing, Data Mining, Knowledge Discovery, Advanced Database Systems, Distributed Processing, Parallel Processing, Algorithms, Systems Architecture. She has authored two books and many research papers. Biography got published in the millennium edition of 'Who's Who in the World – 2000' of Marquis Who's Who, USA an International Magazine and in the International Magazine 'Who's Who in Science and Engineering – 2005". She serves many international societies and research committees.